Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 100 (2015) 1348 – 1353

Procedia Engineering

www.elsevier.com/locate/procedia

25th DAAAM International Symposium on Intelligent Manufacturing and Automation, DAAAM 2014

Using Probabilistic Models for Missing Data Prediction in Network Industries Performance Measurement Systems

Kristjan Kuhi*, Kati Kõrbe Kaare, Ott Koppel

Department of Logistics and Transport, Tallinn University of Technology, Akadeemia 15A, 12616 Tallinn, Estonia

Abstract

The vast development of information and communication technologies has created new possibilities to acquire and analyze data to take performance measurement systems to next level. Most commonly performance measurement has been known as a financial management tool. Sophisticated new technologies have made it possible to collect continuous real-time data and enabled to start designing and implementing nonfinancial performance measurement systems. Most network industries are undertakings of dominant position and therefore subjects to strict supervision. For the authorities to fulfill their regulatory functions, precise monitoring and systemized feedback on the performance of network industries is essential. The problem lies in non-complete data in terms of missing, faulty or delayed values which might lead to incorrect management decisions. The objective of this paper is to explore the use of mathematical models for missing data prediction in performance measurement systems. Applying deterministic models hide the uncertainty of the value state therefore with higher likelihood false diagnoses occur. Authors propose probabilistic models because likelihood based methods for missing data calculation are able to take into account different parameters and time aspect in a single model to convey more trustworthy estimates in performance measurement systems than traditional methods.

Peer-review under responsibility of DAAAM International Vienna

Keywords: industrial engineering; performance measurement; data collection; probabilistic graphical models

1. Introduction

Network industries can be defined as entities where the institution or its product consists of many interconnected

* Corresponding author. Tel.: + 372-51-22-953.
E-mail address: kristjan.kuhi@gmail.com

Peer-review under responsibility of DAAAM International Vienna
doi:10.1016/j.proeng.2015.01.502

nodes and where the connections among the nodes define the character of commerce in the industry [1]. A node in this context can be an institution, a unit of an institution or its product [2]. Examples of network industries are: transportation networks, postal services, electrical power networks, telecommunication networks etc.

When observed parameters of a system or a process are not delivered to processing entity (e.g. to performance measurement system, PeMS) in time, we consider the data missing. The problem of missing data afflicts a variety of application areas in network industries [3,4]. The datasets available to build models are often characterized by missing values, due to various causes such as sensor faults, problems of not reacting experiments, not recovering work situations, transferring data to digital systems [5].

PeMS are reliant on existence of data. Performance cannot be managed and controlled when the feedback cycle does not function properly or delivers faulty indications of the system state. In another hand missing data is everyday problem in statistics. Little and Rubin [6] explain that the mechanisms that lead to missing data are grouped into three distinct groups: missing completely at random (MCAR) – the absence of a data element is not associated with any other value in the data set, observed or missing; missing at random (MAR) – the absence of a values depend only on the observed values in the data set, not on the values that are missing; not missing at random (NMAR) – the absence of a value depends on the other missing values in the data set. In this paper we concentrate on the MCAR and MAR mechanisms where the absence of data does not have information value.

By applying this limitation, the methods for coping with missing values can be grouped into three main categories [6]: inference restricted to complete data (missing values are excluded), imputation-based approaches (missing values are filled in manually), and likelihood-based approaches (missing values are predicted).

The objective of this paper is to explore using mathematical models for missing data prediction in performance measurement systems. Constructing structured probabilistic models of the performance indicators (PIs) taking into consideration the surrounding indicator environment enables to find the indicator values with highest likelihood. It assists to fill gaps in data to improve the quality of the performance analysis and management decisions. In this paper authors present how to apply probabilistic graphical models to performance measurement in network industries that provides support for decision making in case of partial measurement data.

2. Hierarchical model for performance measurement

Performance measurement is the use of statistical evidence to determine progress toward specifically defined social or organizational objectives. Performance measurement describes also the feedback or information on activities with respect to meeting strategic objectives. They are used to measure and improve the efficiency and the quality of the production processes, and identify opportunities for progressive improvements in process performance. Most traditional measures overlook key non-financial PIs [7,8].

Performance is a term used in engineering, in economics and in many other areas. It can have a general meaning or a specific meaning. For the latter, and particularly for network industries, performance must be a measurable entity. Performance measurement techniques represent a key element of network industries asset management systems. Data collection for these systems is becoming feasible due to innovative technological advancements. This is essential for assessing the current and future state of specific fields and management efficiency in productivity, cost-effectiveness, environmental protection, preservation of investments and other functions.

According to literature contemporary PeMS should meet the following criteria: support strategic objectives; have an appropriate balance; have a limited number of performance measures; be easily accessible; consist of performance measures that have comprehensible specifications [9]. Other issues that should be considered selecting performance measures that can be used in evaluation includes forecast ability, clarity, usefulness, ability to diagnose problems, temporal effects and relevance [10,11].

Fig. 1 describes the conceptual model for PIs in the example of road industry. Technical parameter (TP) is measurable or observable environment characteristic. TP has value which varies over time. Uniform PIs permit an evaluation of the effects of different network design and maintenance strategies, but they can also be a basis for predicting network industries performance and for improving old and developing new prediction models. PIs are defined for different types of pavement structures and road categories. In a first step several single PIs describing the characteristic of the road pavement condition are assessed.

The next step is the grouping of these single PIs or indexes into representative combined performance indexes (CPIs) as:

- functional PIs (demands made on road pavements by road users);
- structural PIs (structural demands to be met by the road pavement);
- environmental PIs (demands made on road pavements from an environmental perspective) [10].

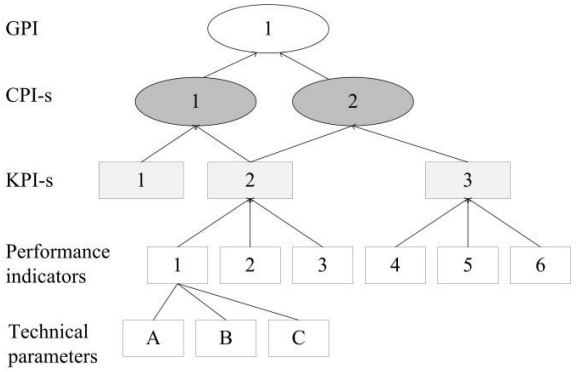

Fig. 1. Conceptual model for road PIs [12-15].

Finally, based on the CPIs a generic PI (GPI) is defined for describing the overall condition of the road network, which can be used for general optimization procedures [13]. For designing PeMS data should be collected about various structural PIs about the road network and about characteristics having an impact on the deterioration of the road. Consequently the performance index is a management tool that allows multiple sets of information to be compiled into an overall measure [16]. Similar trees or graphs are common in other network industries to understand the performance of the institution.

3. Method for missing data prediction

The performance of a system is often uncertain because the observations about the system features are partial, noisy or insignificant features are observed given the exact time moment. When using deterministic models, uncertainty of the system state is not visible, thus applying deterministic models are more likely to give false diagnosis [14].

Missing data can be approximated by multitude of different methods. Numerical analysis methods such as interpolation, extrapolation, time series modeling or likelihood based probabilistic modeling are the most common [6,7]. In this paper authors' explore Bayesian belief networks (BN) as a method that take into account expert knowledge, network element similarities and historical changes in times series in a single model, to calculate the probabilities of missing values.

A BN is a directed acyclic graph representing the joint probability distribution of all variables in a domain. The topology of the network conveys direct information about the dependency between the variables. In particular, it represents which variables are conditionally independent given another variable [7].

Given the knowledge represented as a BN, it can be used to reason about the consequences of specific input data, by what is called probabilistic reasoning. This consists of assigning a value to the input variables, and propagating their effect through the network to update the probability of the hypothesis variables. The updating of the certainty measures is consistent with probability theory, based on the application of Bayesian calculus and the dependencies represented in the network. Several algorithms have been proposed for this probability propagation. BNs can use historical data to acquire knowledge and assimilate domain experts' input [6-7,17].

Dynamic Bayesian Networks (DBN) is an attempt to add temporal dimension into the BN model [18]. Authors presume a time series dataset. Fig. 2 illustrates the probabilistic model. Variable $X=\{x_0, x_1, \ldots, x_t\}$ represents the

variable to be estimated, variable Y={y_0, y_1, ... ,y_t} represent piece of Bayesian network corresponding to all the related variables to X. X_{t+1} represents the value of variable X at the time t+1, and X_{t-1} represents the value of variable X at the time t-1.

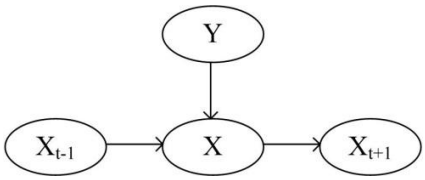

Fig. 2. Dynamic probabilistic model for data estimation [7].

In such system the joint probability factors out with the equation (1) [6].

$$P(X,Y) = P(x_0)\prod_{t=1}^{T-1} P(x_t \mid x_{t-1})\prod_{t=1}^{T-1} P(y_t \mid x_t) \tag{1}$$

If we divide $P(x_t|x_{t-1})$ and $P(y_t|x_t)$ to deterministic and stochastic components assuming linear deterministic part and Gaussian noise part then we can write it down like equations (2) and (3).

$$X_t = AX_{t-1} + w_t \tag{2}$$

$$Y_t = CX_t + v_t \tag{3}$$

Where A is the state transition array and C is the observation array, and v_t, w_t are random noise vectors. This model represents a dynamic model which provides accurate information for estimating the variable in two senses: firstly, using related information identified by experts in the domain; secondly, using information of the previous and incoming values. This information includes the change rate of the variable according to the history of the signal.

DBNs represent the state of the system at different points in time, but do not represent time explicitly. As a consequence, it is very difficult to query a DBN for a distribution over the time at which a particular event takes place. Nodelman, Shelton and Koller [19] have presented an algorithm for approximate inference which takes advantage of the structure within the process over continuous time BN.

4. Applying DBN to performance measurement

Prediction in BN starts with some prior knowledge about the model structure: a set of edges and nodes in belief network. Performance index graphs are common and structured way to represent the expert knowledge concerning dependencies between measurable TPs and PIs (see e.g. [2]). Combining those graphs with time series data, we can build an initial dependency graph, which fragment is represented on Fig. 3, of PIs and CPIs to use for missing data estimation. The exact inference in DBNs to estimate continues values is difficult if not possible to achieve. The method can be applied to estimate discrete PI values since PIs are usually discrete by nature and the quantization algorithms has been proposed by the standards. This allows simplifying and looking the model as linear Gaussian state-space model.

If we take the example of road networks (see [14] for further details) performance monitoring, then R represents rutting PI, C-CPI is comfort index, S-CPI is safety index. S_1 and S_2 are similar road sections. In similar way all the dependencies are captured on a graph to build initial models. Some of the currently missing values can be derived from other measurements and similar historical cases. Some can be approximated and still used in the performance calculus giving indication to the end user of the probability of correctness [20]. In this way we can build a model in each network industries having performance measurement standard in place. This initial model gives us prior

probability distribution over model structure. Enriching it with data will lead to posterior probability of the parameters.

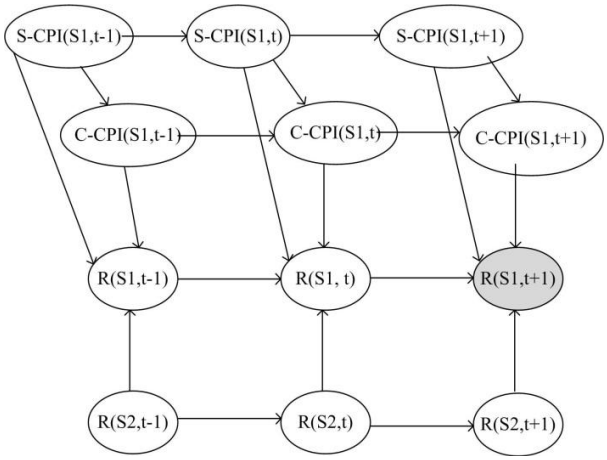

Fig. 3. Fragment of initial dependency graph.

If E={e} is the collection of evidence containing observations about S-CPI, C-CPI, R, S_1 and S_2 over {1:t} then we can build equation 4 to solve the first step prediction problem for time slice t+1.

$$P(R_{t+1;S1} \mid e_{1:t}) = \sum_{R_t} P(R_{t+1} \mid X_t) P(X_t \mid e_{1:t}) \tag{4}$$

A technique of handling estimation issue in this type of networks is the Expectation-Maximization (EM) algorithm. This algorithm makes use of iterative update steps, each time regulating the parameters so as they maximize the expected logarithmic-likelihood, where the expectation is taken with respect to the previous value of the parameters. This algorithm is guaranteed to converge to a locally optimal solution whenever the missing or hidden data are missing at random, i.e. without there being any dependence of the missing data pattern on the values the variables would have had, had they been observed [6].

5. Conclusions

Reliable and accurate predictions of performance can save significant amounts of public resources through better planning, maintenance, and rehabilitation activities. Non-complete data in terms of absent or defective values and deferred timing leads to incorrect decisions. The objective of this paper was to explore using mathematical models for missing data prediction in performance measurement systems

Authors propose an approach to compute missing data using probabilistic graphical models in combination with expert knowledge accumulated in performance measurement best practices to eliminate gaps in network industries performance data. Applying deterministic models hide the uncertainty of the value state therefore with higher likelihood false diagnoses occur. Using probabilistic models is suggested because likelihood based methods for missing data calculation are able to take into account different parameters and time aspect in a single model to convey more trustworthy estimates in PeMSs than traditional methods.

Main conclusions are the following:

- Likelihood based methods for missing data calculation are able to take into account different parameters and time aspect in a single model to convey more trustworthy estimates than traditional methods.

- DBN can easily be created from expert knowledge (in the form of standards and/or best practices), network element similarities and historical changes in times series and provide algorithm for approximate inference on PI level over continuous time.
- This allows filling critical gaps in performance monitoring data together with estimation value likelihood. It will enable making missing data, estimations and their trustworthiness visible to the decision makers.
- The decisions and historical drilldown do not depend on full datasets in each moment of time. Network industries can benefit from the visibility of their performance data.

The conclusions are in line with the previous research. Future research includes practical comparison on the effectiveness of such approach in comparison to unsupervised learning and other possible likelihood based estimation algorithms. Furthermore the research on excluded NMAR mechanism needs to be performed. Testing the model in real-life scenarios and practical optimization of the algorithm can be taken as further steps.

References

[1] H. Göttinger, Economies of Network Industries, Routledge, London, 2003.
[2] K. Kuhi, K. Kõrbe Kaare, O. Koppel, Performance measurement in network industries: example of power distribution and road networks, in: T. Otto (Ed.), Proc. 9th Int. Conf. DAAAM Baltic Industrial Engineering, TUT Press, Tallinn, 2014.
[3] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, F. Li, A tensor-based method for missing traffic data completion, Transportation Research Part C, 28(2013) 15-27.
[4] C. Stenström, A. Parida, Measuring performance of linear assets considering their spatial extension, J. Quality in Maintenance Engineering, 3(2014) 276-289.
[5] P. Baraldi, F. Di Maio, D. Genini, E. Zio, Reconstruction of missing data in multidimensional time series by fuzzy similarity, Applied Soft Computing, 1(2015) 1-9.
[6] R.J. Little, D.B. Rubin, Statistical analysis with missing data, John Wiley & Sons, Hoboken, 2002.
[7] P.H. Ibargüengoytia, U.A. Garcıa, J. Herrera-Vega, P. Hernandez-Leal, E.F. Morales, L.E. Sucar, et al., On the Estimation of Missing Data in Incomplete Databases: Autoregressive Bayesian Networks, in: R. Ege, L. Koszalka (Eds.), 8th Int. Conf. Systems ICONS 2013, Seville, 2013.
[8] T. Wegelius-Lehtonen, Performance measurement in construction logistics, Int. J. Production Economics, 69(2001) 107-116.
[9] S. Tangen, Performance management: from philosophy to practice, Int. J. Productivity and Performance Management, 8(2004) 726-737.
[10] Performance Measures for Road Networks: A Survey of Canadian Use, Transportation Association of Canada, Ottawa, 2006.
[11] P. Vazan, M. Kebisek, P. Tanuska, D. Jurovata, The Data Warehouse suggestion for a production system, in: B. Katalinic (Ed.), Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, DAAAM International, Vienna, 2011.
[12] K.K. Kaare, O. Koppel, Performance measurement data as an input in national transportation policy, in: The XXVIII Int. Baltic Road Conf., Baltic Road Association, Vilnius, 2013.
[13] D. Osborne, T. Gaebler, Reinventing Government, Addison-Wesley, Boston, 1992.
[14] K. Kõrbe Kaare, Performance Measurement of a Road Network: A Conceptual and Technological Approach for Estonia, TUT Press, Tallinn, 2013.
[15] M.A. Ismail, R. Sadiq, H.R. Soleymani, S. Tesfamariam, Developing a road performance index using a Bayesian belief network model, J. Franklin Institute, 348(2011) 2539-2555.
[16] How to Measure Performance: A Handbook of Techniques and Tools, U.S. Department of Energy, Washington, 1995.
[17] D. Smith, G. Timms, P. De Souza, C. D'Este, A Bayesian Framework for the Automated Online Assessment of Sensor Data Quality, Sensors (Basel), 7(2012) 9476-9501.
[18] T. Dean, K. Kanazawa, A Model for Reasoning About Persistence and Causation, Computational Intelligence, 2(1989) 142-150.
[19] U. Nodelman, C.R. Shelton, D. Koller, Continuous Time Bayesian Networks, in: Proc. 18th conf. Uncertainty in artificial intelligence, Morgan Kaufmann, San Francisco, 2002.
[20] J.H. Koskinen, G.L. Robins, P. Wang, P.E. Pattison, Bayesian analysis for partially observed network data, missing ties, attributes and actors, Social Networks, 35(2013) 514-527.
[21] Performance Indicators for the Road Sector. Summary of the field tests, OECD Publications, Paris, 2001.